Indexing and Editing metadata for documentary films on line: the ECHO Digital Library

Giuseppe Amato, Claudio Gennaro, Pasquale Savino

amato/gennaro/savino@iei.pi.cnr.it

I.E.I. – C.N.R.

Via G. Moruzzi, 1 – 56124 Pisa – Italy

Abstract

Wide access to large information collections is of great potential importance in many aspects - economic, environmental, health, cultural, social, etc. - of everyday life. Historical video documentaries hold by national audiovisual archives, constitute one of the most precious - from a historical and cultural viewpoint - and less accessible cultural information.

This paper presents the ECHO (European CHronicles On line) Digital Library which aims at providing a reusable software infrastructure and new metadata models for old historical documentary films in order to support the development of interoperable audiovisual digital libraries. Particular emphasis is given to the Metadata Editor which is designed to be extendible on the basis of the XML schema of the metadata model.

The ECHO Digital Library is being developed within the ECHO Project funded by the European Commission.

KEYWORDS

Audio/Video, Digital Library, Video Documentary, Information Retrieval, Metadata, Metadata Editor, XML schema

1. Introduction

Wide access to large information collections is of great potential importance in many aspects - economic, environmental, health, cultural, social, etc. - of everyday life. However, limitations in information and communication technologies have, so far, prevented the average person from taking much advantage of existing resources. Humanity, in its continuous evolution, has accumulated an enormous quantity of information, knowledge, experience, art treasures, etc. One only has to think of the art treasures contained in our archives, libraries and museums, or of the immense and precious collections of observational data in the areas of space exploration, earth sciences, the environment, medicine, etc., accumulated during the last century. A huge amount of material has also been produced as video material. Many countries have national audiovisual archives, as well as there exist private collections of historical documentaries produced during the twentieth century. Such material is extremely precious from a historical and cultural viewpoint.

The ECHO project [1] aims at developing a Digital Library (DL) service for historical films belonging to large audiovisual archives. Actually being able to see and hear an account of a historical event, filmed in the original context, is very different from reading about it. The ECHO services will allow a user to search and access these documentary film collections. Users will be able, for example, to see an event which is documented in the country of origin and how the same event has been documented in other countries, or to investigate how different countries have documented a particular historical period of their life, etc. One effect of the emerging digital library environment, is that it frees users and collections from geographic constraints. This that we have to work across languages, cultures, international standards, etc. The project involves a number of European institutions holding or managing unique collections of documentary films, dating from the beginning of the century until the seventies. These collections are of great value since they document the different aspects (social, cultural, political, economic) of life in European countries during this period of time. The set of services implemented by ECHO provide users

with access to significant portions of their cultural heritage, which would otherwise be almost inaccessible.

This paper describes the ECHO system, looking at its functionality first and then considering the system architecture and the components used for A/V indexing and retrieval. All these functionality have been defined according to the requirements collected from a user needs analysis performed in the project.

The paper is organized as follows: next sections briefly report the results of the user requirement analysis, while section 3 presents the ECHO system functionality. Section 4 provides a description of the ECHO system architecture; section 5 provides the conclusions and illustrates the future research work.

2. The User Needs

The collection of the user requirements for an audio/video digital library is quite complex: indeed, the system should be usable by a large variety of users, which may have different needs. For this reason the people interviewed were selected from the following categories, all composed of potential ECHO users: (i) educational environment, (ii) researchers studying contemporary or film history, (iii) archivists, (iv) producers of A/V products, and (v) people marketing A/V products.

There are several ways for collecting user requirements; amongst all the possibilities, we decided to carefully select the users to be interviewed, to prepare a quite detailed questionnaire and to interview them in a face-to-face meeting. The questionnaire included questions on different areas, such as (a) how the user want the data entry to be managed (this section was mainly devoted to archivists), (b) what are the indexing features the user wants supported, (c) what are the required retrieval functionality, (d) how the query results should be presented, (e) what are the requirements for the reuse of the retrieved material, (f) what is the billing and accounting support needed, (g) which kind of material should be included in the digital library.

The analysis of the interviewees was quite extensive and the results covered many aspects of the ECHO system. However, in this paper we will briefly review only the most relevant topics; the interested reader

may access the full project deliverable [2]. The requirements have been subdivided into the different areas of interest.

- **Data entry management.** The documentalist consider relevant to have the possibility for an automatic insertion of A/V documentaries, plus the support of a metadata editor that allows one to insert other, more specific metadata and to correct the manually inserted metadata.
- **Indexing functionality.** Indexing should be based on manually extracted features (such as speech transcripts, image features, faces, particular objects present in the video), as well as on a powerful metadata description that allows one to provide a detailed view of the documents.
- **Retrieval functionality.** The retrieval should be based on all metadata used during the indexing, supporting free text and structured searching. Exact as well partial match retrieval (with ranking of the retrieved documents in decreasing relevance order with the query) is required. The access to single or multiple distributed archives is required. A support for cross-language retrieval is needed: the user should be able to formulate the query in one language and to retrieve documents that have been indexed using a different language.
- **Result's presentation.** Query results should be presented in decreasing relevance order with the query. The presentation of the results should be personalized: the user should be able to specify the items of the metadata description he prefers to be presented. Furthermore, a visual abstract (having a length specified by the user) should be used for a fast browsing of query results.
- **Reuse of the Digital Library content.** It should be supported the export of the entire video, of parts of the video, of the abstract, and of the video metadata.

3 ECHO System Functionality

The requirements specified so far, led to the definition of a number of functionality that are briefly described in the following. A particularly relevant requirement consists in the possibility to support the *interoperability* of distributed, heterogeneous digital collections and services. Achieving interoperability

among digital libraries is facilitated by conformance to an open architecture as well as agreement on items such as formats, data types, and metadata conventions.

In order to render a digital library of this type feasible, the project has to solve the numerous technical problems that currently bar the inclusion of film information in the digital environment. The aim is to make the film collections available to as broad as possible range of users. To achieve this goal, the following activities have been performed and the following modules have been included into the system:

Define an A/V metadata model.

An important aspect of the project consists in the addition of a layer of metadata to the film archives. Metadata elements as presently defined do not describe film data well. A semi-automatic process was designed by which existing local catalogue records can be integrated with metadata elements, automatically extracted during the indexing/segmentation of the film material, into a common description, i.e., the common metadata model.

Design a multilingual retrieval user interface.

Local site interfaces were implemented in the local languages; however, a common user interface in English will also be maintained on the project Web-site for external access. Online cross-language search tools were included. Cross-language interrogation is enabled by means of the employment of standard metadata formats, and via mechanisms which provide a mapping between the descriptive languages used by each partner.

Provide an intelligent access to digital films

The utility of the digital film library can be judged by the ability of the users to retrieve information they need easily and efficiently. The project provides content-based searching and film-sequence retrieval. As the content is conveyed in both narrative (text and speech) and the image, a collaborative interaction of image, speech and language technology will be adopted in order to search the diverse film collections with satisfactory effectiveness. Speech recognisers for different languages are being integrated into the system.

Create visual summaries

The project is developing techniques to produce visual summaries. The aim is to capture the content and structure of the underlying documentary film in a brief visual abstracting process. The summary consists of a sequence of moving images, much shorter than the original film, but preserving the essence of the original message. It should provide a good overview of the entire film documentary.

Protect intellectual property rights; providing security, privacy, and accounting

In order to make a digital library of films possible, the copyright owners must be assured that their property will be properly protected and that its use will be measured in order to ensure them appropriate compensation. The ECHO system, therefore, includes mechanisms which support access control, authentication, security, privacy, and billing.

4 The ECHO system

The ECHO system assists the user during the indexing and retrieval of A/V documentaries.

The indexing is semi-automatic. Using a high-quality speech recogniser, the sound track of each video source is converted to a textual transcript, with varying word error rates. The transcript is then stored in a full-text information retrieval system. Multiple speech recognition modules, for different European languages will be included. Likewise, video and image analysis techniques are used for segmenting video sequences by automatically locating boundaries of shots, scenes, and conversations. Metadata is then manually associated with film documentaries in order to complete their classification.

Search and retrieval via desktop computer and wide area networks is performed by expressing queries on the audio transcript, on metadata or by image similarity retrieval. Retrieved documentries or their abstract, are then presented to the user. By the collaborative interaction of image, speech and natural language understanding technology, the system compensates for problems of interpretation and search in the error-full and ambiguous data sets. Exploration of the ECHO library is based on these same techniques, allowing for spoken or typed natural language access to the information space.

4.1 Overall system architecture

The architecture of the ECHO system (Figure 1) is composed of three main components: *client interface*, *automatic processor* and *middlewere*. The client interface is the component directly employed by the users to interact with the system. The automatic processor component analyses multimedia documents, to automatically extract metadata. The middlewere component manages accesses to data stored in the video database and metadata database, on behalf of the other two components. In the following, we will give a more detailed description of each of these components.

4.1.1 Client interface

The Client interface is composed of four main modules, related to the corresponding activities that can be carried out by users of the system. The *metadata editor* allows users to manually edit and review metadata associated with multimedia documents. The user can either edit automatically generated metadata, as for

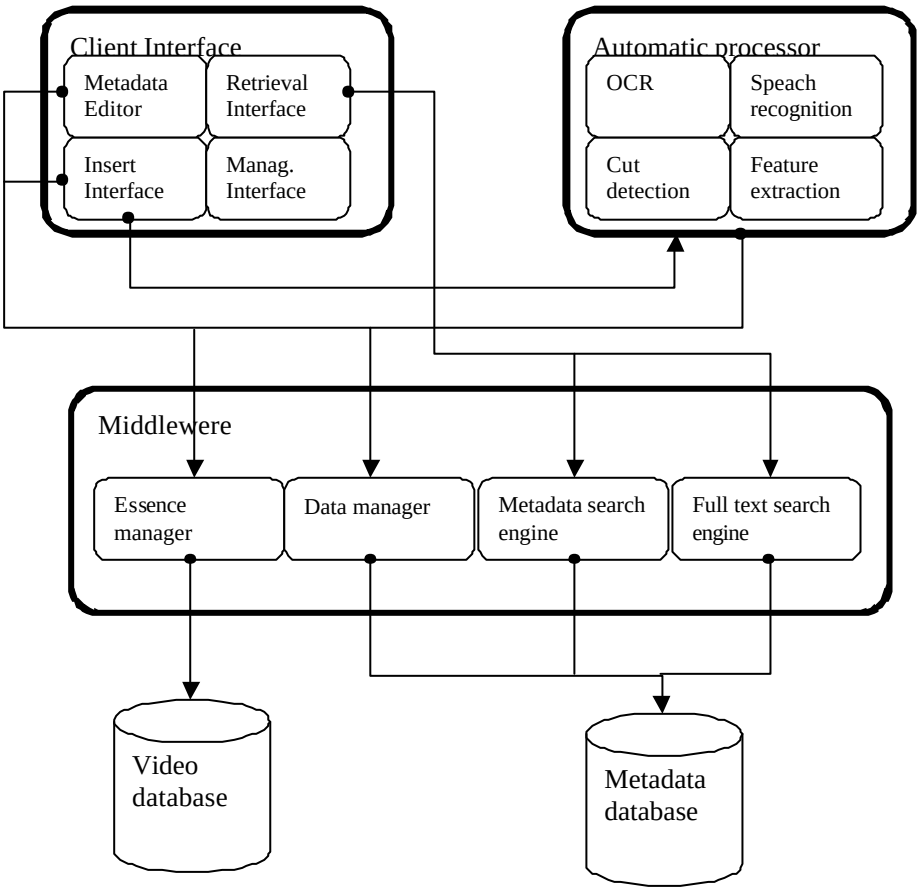

Figure 1: ECHO system architecture.

instance scene boundaries, or he can add additional metadata manually. The *insert interface* is used when new documents are inserted. This module interacts with the metadata editor and the automatic processor components that automatically analyse the documents being inserted. The *retrieval interface* is used to search the system for interesting documents. Various possibilities are offered by this interface. Users can retrieve documents by performing full text retrieval, on the transcript or descriptions associated with documents, or selecting specific field of the metadata structure. Finally, the *management interface* can be used to configure and fine-tune the system.

4.1.2 Automatic processor

The Automatic processor is composed of four main modules, each one dedicated to a different automatic processing technique. The *OCR* module recognises textual video captions. The *speech recognition* module is able to generate a transcript in correspondence of an audio or audio/video document. The generated transcript is indexed and the corresponding document can be retrieved by performing full text retrieval. The *cut detection* module analyses a video document and automatically identifies scene changes. In this way, metadata can be associated with specific portions of the document, instead of the whole document. Finally, the *feature extraction* module analyses multimedia document in order to extract physical properties, which can be used to perform similarity retrieval. Typical features extracted are colour distribution, texture, shapes, and motion vectors [3].

4.1.3 The middlewere

This component manages the accesses to the underlying databases: the video database, that physically stores video documents managed by the system, and the metadata database, where all metadata associated with the documents are stored. The middlewere component is constituted of four modules. The *essence manager*, that handles the access to the video server, when other system modules, e.g. the metadata editor or the automatic processor, need to access to the video server. The *data manager*, that handles the access to the metadata database. Main operations supported are *read, insert, delete*, and *modify* metadata fields. The *metadata search engine* supports document search based on the full metadata content. The *full text*

search engine can be used to search for documents by using textual parts of the metadata. For instance, the descriptions and the transcripts associated with documents are used to perform this type of search. The result of the two search engines can be combined to allow users to perform an integrated complex search.

4.2 Editing Metadata

A fundamental aspect of the ECHO project is the metadata model [4] used for representing the audiovisual contents of the archive. The proposed model is based on the IFLA model, a general conceptual framework used to describe heterogeneous digital media resources [5]. This metadata model is composed of four levels describing different aspects of intellectual or artistic endeavour: *work*, *expression*, *manifestation*, and *item*. The entities of the model are organized in a structure that reflects the hierarchical order of the entities from the top level (*work*) to the bottom (*item*). Figure 2 shows a schematic representation of the ECHO Metadata Model [4].

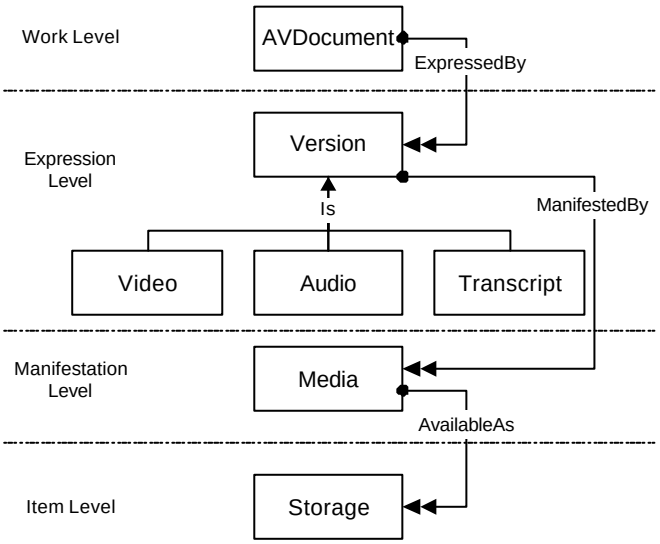

Figure 2: schematic representation of the Echo Metadata Model.

The AVDocument entity is the most abstract one; it provides the general intellectual or artistic view of the document. For instance, let us suppose we want to describe a document about the Berlin Olympic Games

in 1936. An AVDocument object will represent the abstract idea of the documentary film on the Games. A number of objects, of the abstract entity Version, could represent different language version of the same film, e.g., versions in Italian or in German. The Italian version could have three different objects that are specializations of the entity Version: a Video object, a Transcript object and an Audio object. Moreover, Version objects can be also part of other Version objects. For example, the Video object representing the Italian version of the whole documentary film can have other Video objects representing specific scenes of the film.

However, the Version entity does not represent any specific implementation of the film. This aspect can be represented by means of the manifestation level. For instance, a Media object could represent a digital realization of the document in MPEG format. More than one manifestation of the same Version, e.g. MPEG, AVI, etc., may exist.

Nevertheless, the Media object does not refer to any physical implementation. For instance, the MPEG version of the Italian version of the Games can be available on different physical supports, each one represented by a different Storage object (e.g., videoserver, DVD, etc).

Since the metadata model is relatively complex, the design of the metadata editor is of primary importance. The editor is intended to be used by the cataloguers of the archive, that insert new audiovisual documents and that specify the metadata of the documents. The typical cataloguer workflow is the following:

1. A new audiovisual document is digitalized or transformed from one digital format into another;
2. The document is archived by the system in the videoserver;
3. The document is processed for automatic indexing (extraction of scene cuts, speech recognition, etc.);
4. When the automatic indexing has been completed, the user is informed by the system and the manual indexing can start;
5. The user typically edits the textual description for typos or factual content, reviews or sets values of the metadata fields, adjusts the bounds of the document segments, removes unwonted segments and merges multiple documents. This phase is usually performed starting from the top level of the model

(the AVDocument), and continuing by modifying/editing the lower-level objects connected to the AVDocument (i.e., Version, Media and Storage objects).

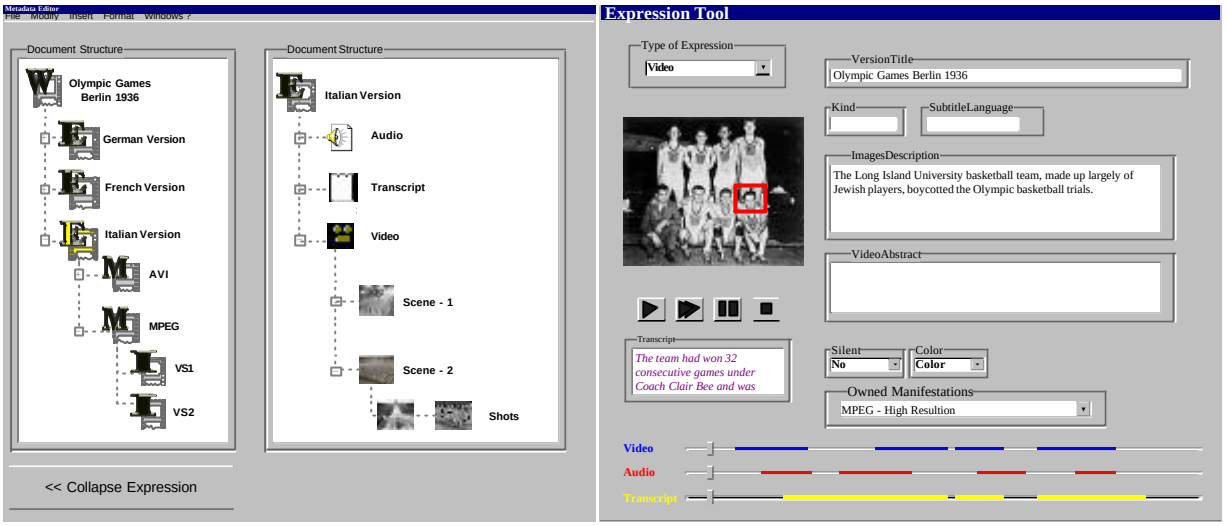

Figure 3: A screenshot of the Metadata Editor. Document structure (left) and Expression Tool (right).

The interface of the editor is designed in such a way that it is possible to browse the tree structure of an audiovideo document. Figure 3 shows a screenshot of the interface: the window on left side displays a document like a folder navigation tool. On the top level of the tree, there is an icon representing an AVDocument object (the work of the "Olympic Games on 1936" in our example). Connected to the work object the editor presents the three main Versions that belong to the AVDocument. Moreover, selecting an icon representing a Version (the Italian Version in Figure), it is possible to see the Media instances of the Version and, hence, the corresponding Storage objects.

The navigation tool on the left side of the window shows only the main expressions belonging to the documents (i.e., the expression which correspond to the whole audiovideo document). The editor allows to browse a single Version one at a time by using a second frame on the right side of the window. In this way it is possible to see the possible Video, Audio and Transcript Versions (at least one of them must exist) of the document and, for each Version, to browse the video segmentations in scenes, shots, etc.

By clicking on the icon corresponding to a metadata object, it is possible to modify, in a separated window, the metadata fields of the object. A particular attention has been paid to the expression window

design, i.e, the Expression Tool. Figure 3 gives an example of the Expression tool interface. Besides the textual fields, the Expression Tool allows the access to the metadata relative to the video segmentation, and allows one to modify them. More precisely, the user can view the video, hear the audio and read the transcript. The window shows also an overview of the video segmentation, by means of three slide tools (see the bottom of the Expression Tool window), which represent the video, the audio and the transcript (if any) of the whole expression. These slides are subdivided in partition that represent the media segmentation. By selecting a segment, the Expression Tool shows the Version corresponding to the subpart of the media (for instance, a scene or a shot).

4.2.1 The editor architecture

As has previously been shown in section 4.1, the metadata editor communicates with the middlewere of the Echo system in order to obtain the metadata of the A/V documents from the database. In particular, once the user has found a relevant document, by means of the video retrieval tool (see section 4.2.2), the URI (Uniform Resource Identifier) of the document is obtained and passed to the metadata editor. This URI is sent to the datamanager which returns the document metadata. The format chosen for exchanging document metadata are is XML. The metadata editor is not hard wired with a particular metadata attributes set, indeed the metadata schema is defined in the W3C XML Schema Definition (XSL) and is used by the editor as configuration file for the metadata model. The advantage of this choice is that it is possible to add/remove fields in the schema of the metadata of the audiovisual document (see Figure 4). This is achieved by giving the editor the ability of recognizing a subset of the types available for the XSL schemas; which are: **`xsd:string`**, **`xsd:bool`**, **`xsd:date`**, **`xsd:float`**, **`xsd:integer`** and **`SetOfString`** (the tag "**`xsd:`**" means that the type is built in to XML schema). The special type **`SetOfString`** has been introduced in order to allow the schema designer to have multiple fields.

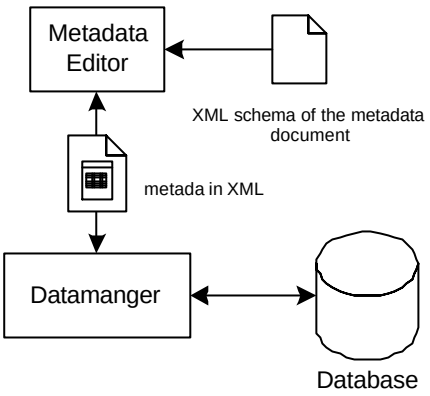

Figure 4: The architecture of the metadata editor

In the following an example a subset of the AVDocument entity schema is given:

```
<?xml version="1.0" encoding="UTF-8"?>

<xsd:schema xmlns:xsd="http://www.w3.org/2000/10/XMLSchema" elementFormDefault="qualified">

   <xsd:element name="AVDocument">
      <xsd:annotation>
         <xsd:documentation>work level entity</xsd:documentation>
      </xsd:annotation>
         <xsd:complexType>
            <xsd:sequence>

               <xsd:element name="Title" type="xsd:string">
                  <xsd:annotation>
                     <xsd:documentation>Original title if known otherwise
                     assigned</xsd:documentation>
                  </xsd:annotation>
               </xsd:element>

               <xsd:element name="Genre" type="xsd:string">
                  <xsd:annotation>
                     <xsd:documentation>Genre of the Document</xsd:documentation>
                  </xsd:annotation>
               </xsd:element>

               ...
               other fields...
               ...

            </xsd:sequence>
      </xsd:complexType>
   </xsd:element>

   <xsd:complexType name="SetOfStrings">
      <xsd:sequence>
         <xsd:element name="string_item" minOccurs="0" maxOccurs="unbounded"/>
      </xsd:sequence>
   </xsd:complexType>
   <xsd:simpleType name="string_item">
      <xsd:restriction base="xsd:string"/>
   </xsd:simpleType>

</xsd:schema>
```

Inside the tag **<sequence>** (see the XSL documentation for more details [6]) a metadata field for the AVDocument entity can be defined by means of the tag **<xsd:element>**. The tag **<xsd:annotation>** it is useful for the reader in order to better understand the meaning of the field and it is used also by the editor as a ToolTip when the mouse pointer is moved over the form's input control related to field.

As an example, suppose a new field, that indicates the date of the A/V Document, (called *ProductionDate*) is needed. It is sufficient to put a new element as the following:

```
<xsd:element name="ProductionDate" type="xsd:date">
   <xsd:annotation>
      <xsd:documentation>Date of the production</xsd:documentation>
   </xsd:annotation>
</xsd:element>
```

Eventually, the instance of a A/V Document based on the presented schema could be the following:

```
<?xml version="1.0" encoding="UTF-8"?>

<AVDocument xmlns:xsi="http://www.w3.org/2000/10/XMLSchema-instance"
xsi:noNamespaceSchemaLocation="C:\Metadata Editor\XML\Schema\AVDoc.xsd">
   <Title> Olympic Games on 1936</Title>
   <Genre>Documentary</Genre>
   <Description>Documentary on 193 Berlin Olympic Games</Description>
   <Person_names>
      <string_item>Jesse Owens</string_item>
      <string_item>Hendrika Mastenbroek</string_item>
   </Person_names>

   ...

</AVDocument>
```

4.2.2 The Video Retrieval Tool

The video retrieval tool allows the user of the archive to search for the audiovideo documents in the database. The interface of the system offers three types of searches: a monolingual free text search, a cross language search and a boolean structured search. Figure 5 shows a simplified version of the retrieval tool interface. On the left side of the windows, three search forms are displayed and on the right side, the documents retrieved are listed.

The monolingual free text search looks for documents whose metadata contain the words specified in the form according to the type of search specified (i.e, "any word" or "all words").

The cross language search allows one to look for documents over three specific metadata fields: Keywords, People and Location. These cross language fields can contain only a closed list of words, which are stored in a database table that contains the translations in all four languages.

The structured search looks for documents by matching the metadata fields given in the form. This type of search is accomplished by translating the query in SQL and by submitting it to the database.

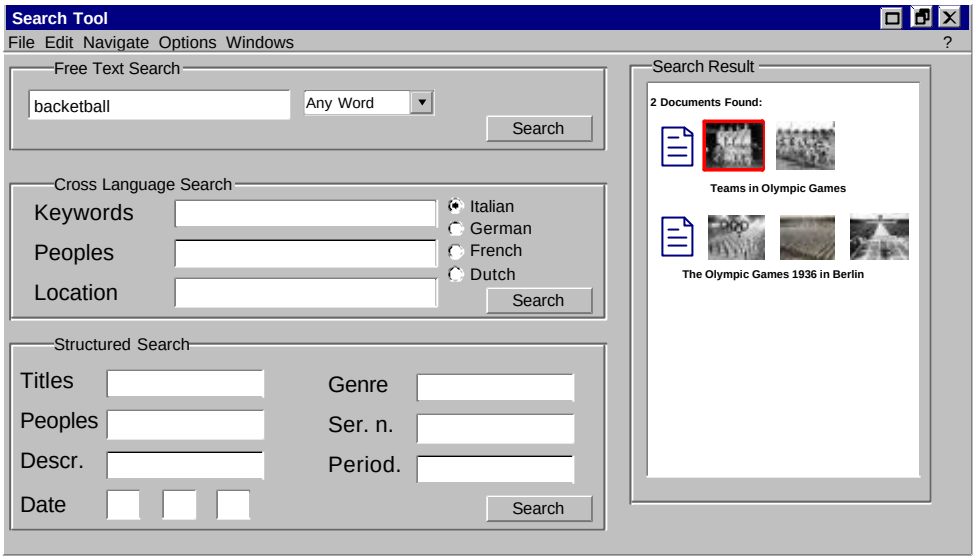

Figure 5: Video Retrieval Tool

When the search is executed, the system shows some keyframes of the video part of the retrieved documents. By selecting a document the user can see the multimedia content and the metadata content by using an interface similar to the interface of the metadata editor.

5 Conclusions

The first ECHO prototype is under development, as well as the metadata editor. The first prototype (*Multiple Language Access to Digital Film Collections*) integrates the speech recognition engines with the infrastructure for content-based indexing. It will demonstrate an open system architecture for digital film libraries with automatically indexed film collections and intelligent access to them on a national language basis. This prototype will be refined and the system functionality will be enhanced in two successive prototypes. The second prototype (*Multilingual Access to Digital Film Collections*) will integrate the

metadata editor, which will be used to index the film collections according to a common metadata model. Index terms, extracted automatically during the indexing/segmentation of the film material (first prototype), will be integrated with local metadata, extracted manually, in a common description (defined by the common metadata model). The second prototype will support the interoperability of the different video collections and content based searching and retrieval. The third prototype (*ECHO Digital Film Library*) will add summarization, authentication, privacy and charging functionalities in order to provide the system with full capabilities.

Acknowledgement

This work has been funded by the 5th F.P. IST Research Program, Project No. 11994, ECHO (European CHronicles On line). We would like to thank the members of the ECHO Consortium for fruitful discussions which inspired the ideas presented in the paper.

References

[1] ECHO Web site: http://pc-erato2.iei.pi.cnr.it/echo

[2] ECHO User Requirement Report, ECHO Project Deliverable D1.2.1, June 2000, http://pc-erato2.iei.pi.cnr.it/echo/workpackages/wp1.html

[3] B. Furht, S.W. Smoliar, H. Zhang, "Video and Image Processing in Multimedia Systems", Kluwer Academic Publishers, 1996. ISBN 0-7923-9604-9

[4] G. Amato, D. Castelli, S. Pisani, "A Metadata Model for Historical Documentary Films", Proc. of the 4th European Conference ECDL 2000, Lisbon, Sept. 2000

[5] K.G. Saur München, "Functional Requirements for Bibliographic Records", Final Report, 1998, http://www.ifla.org/VII/s13/frbr/frbr.pdf

[6] David C. Fallside, "XML Schema Part 0: Primer W3C Proposed Recommendation", 30 March 2001, http://www.w3.org/TR/xmlschema-0/